Optimality regions and fluctuations for Bernoulli last passage models

Article (Published Version)

Math Phys Anal Geom (2018) 21:22
https://doi.org/10.1007/s11040-018-9276-2

Optimality Regions and Fluctuations for Bernoulli Last Passage Models

Nicos Georgiou[1] · **Janosch Ortmann**[2]

Received: 21 March 2018 / Accepted: 14 June 2018

Abstract We study the sequence alignment problem and its independent version, the discrete Hammersley process with an exploration penalty. We obtain rigorous upper bounds for the number of optimality regions in both models near the soft edge. At zero penalty the independent model becomes an exactly solvable model and we identify cases for which the law of the last passage time converges to a Tracy-Widom law.

Keywords Soft edge · Edge results · Optimality regions · Sequence alignment · Discrete Hammersley process · Longest common subsequence · Bernoulli increasing paths · Tracy-Widom distribution · Last passage time · Corner growth models · Flat edge

Mathematics Subject Classification (2010) 60K35

NG was partially supported by the University of Sussex Strategic development Fund (SDF) and by the EPSRC First grant EP/P021409/1: The flat edge in last passage percolation. JO was partially supported by an ISM-CRM fellowship and a Concordia Horizon fellowship

✉ Nicos Georgiou
n.georgiou@sussex.ac.uk
http://www.sussex.ac.uk/profiles/329373

Janosch Ortmann
ortmann.janosch@uqam.ca
http://crm.umontreal.ca/∼ortmann/

[1] Mathematics Department, University of Sussex, Falmer Campus, Brighton BN1 9QH, UK

[2] Département de management et Technologie, Université du Québec à Montréal, Montréal, Canada

Published online: 01 August 2018

1 Introduction

1.1 Directed Growth Models

In this article we study a generalisation of two specific models of directed last passage percolation, namely the *longest common subsequence model* concerning the size of the longest common subsequence between words drawn uniformly from a finite alphabet [8], and an independent version introduced in [40] as an exactly solvable discrete analogue of the Hammersley process [20]. We call the latter the *independent model*.

We study these models near directions for which the corresponding shape function starts developing a flat segment, which is called the *soft edge* of the model. Both models fit in the general framework [14], namely there is:

(i) The random environment $\omega \in \mathbb{R}^{\mathbb{Z}^2}$, whose law we denote by $\mathbb{P}$. Each marginal ω_u should be viewed as a random weight placed on site $u \in \mathbb{Z}^2$.
(ii) A collection Π of admissible paths on $\mathbb{Z}^2$. A path π from u to v is uniquely identified by an ordered sequence of integer sites, so when necessary we write $\pi = \{u = u_0, u_1, \ldots, u_\ell = v\}$. A path π is admissible if and only if its increments $z_k = u_k - u_{k-1}$ are contained in a finite set $\mathcal{R} \subset \mathbb{Z}^2$. For $u, v \in \mathbb{Z}^2$ we denote the set of admissible paths from u to v by $\Pi_{u,v}$. It is a requirement that $\mathbb{P}$ is stationary and ergodic under shifts T_z, $z \in \mathcal{R}$.
(iii) A measurable potential function $V : \mathbb{R}^{\mathbb{Z}^2} \times \mathcal{R}^\ell \to \mathbb{R}$. For the two models under investigation we always have $\ell = 1$ and V is a bounded function, thus satisfying the technical assumptions of [14].

The *point-to-point last passage time* from u to v is the random variable G^V defined by

$$G^V_{u,v} = \max_{\pi \in \Pi_{u,v}} \left\{ \sum_{u_k \in \pi} V\left(T_{u_k}\omega, z_{k+1}\right) \right\}. \tag{1.1}$$

A well studied version of the model is the *corner growth model*, for which $\mathcal{R} = \{e_1, e_2\}$, the coordinates of ω are i.i.d. under $\mathbb{P}$ and the potential V for the corner growth model is defined by

$$V(\omega, z) = \omega_0, \quad z \in \mathcal{R} = \{e_1, e_2\}. \tag{1.2}$$

Whenever we are referring to last passage time under this potential and these admissible steps, we will use T instead of G^V. It is expected that under some regularity assumptions on the moments and continuity of ω_0, the asymptotic behaviour of T (e.g. fluctuation exponents for T and the maximal path, distributional limits, etc) is environment-independent. This is suggested by results available for the two much-studied exactly solvable models when ω_0 is exponentially or geometrically distributed and further evidenced by the general theory in [14–16] and the edge results of [7, 31], as we discuss later.

The main models in this article have set of admissible steps $\mathcal{R} = \{e_1, e_2, e_1 + e_2\}$ and the coordinates of the environment take values in $\{0, 1\}$. Our choice of potential

is a two-parameter family of bounded functions, indexed by two non-negative parameters α and β:

$$V_{\alpha,\beta}(\omega, z) = \begin{cases} \omega_0 - \alpha\,(1-\omega_0) & \text{if} z = e_1 + e_2 \\ -\beta & \text{if} z \in \{e_1, e_2\}\,. \end{cases} \tag{1.3}$$

This particular choice of potential is inspired by a problem which appears in computational molecular biology, computer science and algebraic statistics, as we explain at the end of this introduction. Our strongest results are obtained when $\alpha = \beta = 0$ and the marginals of ω are i.i.d. Bernoulli random variables on $\{0, 1\}$ with parameter $p \in (0, 1)$, because we then obtain a solvable model [39]. This will be referred to as the *independent model*, and the passage time from $(0, 0)$ to (m, n) is denoted $G^{(\alpha,\beta)}_{m,n}$ when both α and β are important. When $\alpha = 0$ we further simplify notation by $G^{(\beta)}_{m,n} = G^{(0,\beta)}_{m,n}$. The special case $\alpha = \beta = 0$ was studied in [5, 13, 40]. Asymptotic results as p tends to zero were obtained in [25].

We consider a rectangle of height n and width $m_n = \frac{n}{p} - xn^a$ for $a \in (0, 1)$ and show that the fluctuations of $G^{(0)}_{m_n,n}$ converge, suitably rescaled, to the Tracy–Widom GUE distribution. The size of the rectangle is not arbitrary. A justification for this option comes by looking at the limiting shape function

$$g_{pp}(t) = \lim_{n\to\infty} \frac{G^{(0)}_{\lfloor nt\rfloor,n}}{n},$$

continuous in t. When $t > \frac{1}{p}$ the function has a *flat edge*: $g_{pp}(t) = 1$.When $p < t < 1/p$, $g_{pp}(t)$ is strictly concave and when $t < p$, g_{pp} has another flat edge, namely $g_{pp}(t) = t$. Fluctuations of $G^{(0)}_{\lfloor nt\rfloor,n}$ are of order $n^{1/3}$ when $t \in \left(p, \frac{1}{p}\right)$, so by looking at the rectangle $m_n \times n$ we study these fluctuations at the onset of the flat edge, but macroscopically we converge to the critical point $t = 1/p$.

1.2 Edge Results

There is a coupling of $G^{(0)}_{\frac{n}{p}-xn^a,n}$ with $T_{n,n^{2a-1}}$, which we describe in Section 4. This mapping was exploited in [13] to obtain the local weak law of large numbers

$$\lim_{n\to\infty} \mathbb{P}\left\{ \left| \frac{n - G^{(0)}_{p^{-1}n-xn^a,n}}{n^{2a-1}} - \frac{(px)^2}{4(1-p)} \right| < \varepsilon \right\} = 1 \tag{1.4}$$

for all $a \in \left(\frac{1}{2}, 1\right)$. We use the same coupling to obtain a distributional limit for the edge. The coupling classifies results for $G^{(0)}_{p^{-1}n-xn^a,n}$ as "edge results". The terminology "edge results" is motivated by the fact that the last passage time T is studied in a thin rectangle, either with dimensions $n \times yn$ and letting $y \to 0$ after sending $n \to \infty$ [31], or with only one macroscopic edge, namely of dimensions $n \times xn^\gamma$ with $\gamma < 1$.

Several results near the edge are universal, in the sense that they do not depend on the particular distribution of the environment. In the sequence we denote the environment for the corner growth model by $\zeta = \{\zeta_u\}_{u\in\mathbb{Z}^2_+}$. An approximation of i.i.d. sums with a Brownian motion [26] was used in [17] to obtain the weak law of large numbers,

$$\frac{T_{n,xn^\gamma} - n\mathbb{E}(\zeta_0)}{\sqrt{\mathrm{Var}(\zeta_0)n^{1+\gamma}}} \Longrightarrow c\sqrt{x}, \quad (n\to\infty),$$

and simulations lead to the conjecture that $c = 2$. The conjecture was proved in [41] via a coupling with an exclusion process and later in [4] using a random matrix approach. A coupling with the Brownian last passage percolation model [4, 36] allow [7] to obtain

$$\frac{T_{n,n^\gamma} - n\mathbb{E}(\zeta_0) - \sqrt{\mathrm{Var}(\zeta_0)n^{1+\gamma}}}{n^{\frac{1}{2}-\frac{\gamma}{6}}\sqrt{\mathrm{Var}(\zeta_0)}} \Longrightarrow W, \quad (n\to\infty), \tag{1.5}$$

where W is has the *Tracy-Widom GUE distribution* [43]: the limiting distribution of the largest eigenvalue of a GUE random matrix. If ζ_0 has exponential moments, (1.5) holds for all $a \in (0, \frac{3}{7})$.

1.3 The Alignment Model

The problem of *sequence alignment* [34, 42] can be cast in this framework. Consider two words $\eta^{\mathrm{x}} = \eta^{\mathrm{x}}_1 \dots \eta^{\mathrm{x}}_m$ and $\eta^{\mathrm{y}} = \eta^{\mathrm{y}}_1 \dots \eta^{\mathrm{y}}_n$ formed from a finite alphabet $\mathcal{A}$. We consider the case where each letter of η^{x} and η^{y} is chosen independently and uniformly at random from $\mathcal{A}$. We are looking for a sequence of elementary operations of minimal cost that transform η^{x} to η^{y}. These operations are:

(1) replace one letter of η^{x} by another, at a cost α
(2) delete a letter of η^{x} or insert another letter, each at a cost of β.

Assign a score of 1 for each match and subtract the costs for replacements, deletions and insertions. Each sequence of operations taking η^{x} to η^{y} is thus assigned a *score* $L^{(\alpha,\beta)}_{m,n}$, also often called the *objective function*. We will also write $L^{(\beta)}_{m,n}$ for $L^{(0,\beta)}_{m,n}$.

A problem arising in molecular biology [1, 21, 35, 37, 44, 46] is to maximise this alignment score. In that context the words η^{x} and η^{y} can be DNA strands (with $\mathcal{A} = \{A, C, G, T\}$), RNA strands ($\mathcal{A} = \{A, C, G, U\}$) or proteins (with $\mathcal{A}$ the set of amino acids that make up a protein), and the elementary operations correspond to mutations. A choice of the parameters α and β corresponds to a judgement on how frequently each type of mutation occurs. The optimal score for an alignment of η^{x} with η^{y} can then be considered a measure of similarity between these words. The question also appears in algebraic statistics [38]: there the objective function is the tropicalisation of a co-ordinate polynomial of a particular hidden Markov model.

The special case $\alpha = \beta = 0$ corresponds to the problem of finding longest common subsequence (LCS) of the words η^{x} and η^{y}, which has been intensively studied by computer scientists [6, 22, 29, 32] and mathematicians [2, 8, 18, 23, 27, 28].

On the other hand, the alignment score $L_{m,n}^{(\alpha,\beta)}$ is the last passage time (1.3) in environment

$$\omega_{ij} = \begin{cases} 1 & \text{if } \eta_i^{\mathrm{x}} = \eta_j^{\mathrm{y}} \\ 0 & \text{otherwise,} \end{cases} \tag{1.6}$$

i.e. the marginals of ω are (correlated) Bernoulli random variables with parameter $|\mathcal{A}|^{-1}$. The model with this choice of environment is referred to as the *alignment model*.

A deletion of a character in η^{x} corresponds to a horizontal step (e_1) in the last passage model, whereas an insertion of a letter into η^{x} corresponds to a vertical step (e_2). Replacing a letter in η^{x} by another corresponds to a diagonal step (e_1+e_2) onto a point (i, j) where $\omega_{ij} = 0$, whereas any letter left alone (i.e. a successful alignment) corresponds to a diagonal step onto a point (i, j) where $\omega_{ij} = 1$. The path in Fig. 1 corresponds to the alignment

$$\eta^{\mathrm{x}} : A\ A\ B\ A\ -B\ A$$
$$\eta^{\mathrm{y}} : -\ A\ B\ A\ A\ B\ A$$

in which the bar under the first A of η^{x} corresponds to deleting the letter A from η^{x} while the bar in η^{x} corresponds to inserting the letter A there. A convenient way to look at this is that the bars, called *gaps*, are used to stretch the two words appropriately so that different matchings are obtained.

1.4 Optimality Regions

Which paths are optimal depends on the choice of parameters α, β. In molecular biology these parameters are often chosen ad hoc and it is not clear that there is a single 'right' choice [44]. An alternative approach is to consider the space $\mathcal{C} = [0, \infty) \times [0, \infty)$ of all possible parameters (α, β) and to analyse how the optimal

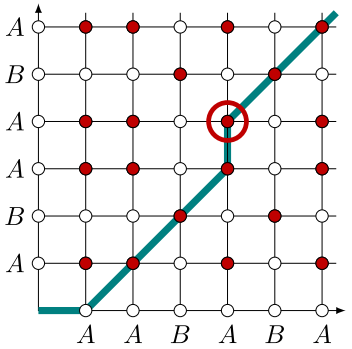

Fig. 1 Environment generated by the two strings $AABABA$ and $ABAABA$. Colored dots correspond to the value 1, white dots to the value 0. The thickset path is a maximal path in this environment, from $(0, 0)$ to $(6, 6)$ with minimal number of vertical or horizontal steps (just 2 in this case). When $\alpha = 0$, the illustrated path has score $5 - 2\beta$ since the environment only contributes to the weights if collected by a diagonal step. The score coincides with the last passage time for $\beta \leq 1/2$. For $\alpha = 0$ and $\beta > 1/2$ the main diagonal is optimal, with score equal to 4. These are the only two optimal paths, so there are two optimality regions

paths change as (α, β) varies. A maximal subset of $\mathcal{C}$ on which the set of optimal paths does not change is called an *optimality region* of $\mathcal{C}$. The shape of optimality regions in $\mathcal{C}$ are semi-infinite cones bounded by the coordinate axes and by lines of the form $\beta = c+\alpha(c+1/2)$ for certain values of c. So it suffices to study the number of regions with one parameter fixed; we will set $\alpha = 0$.

Denote the number of optimality regions in this model by $R^{(\mathrm{al})}_{m,n}$. Naturally the (expected) number of optimality regions attracted a lot of interest both theoretically [12, 19, 45] and in applications [10, 24, 30, 33]. The current conjecture [11, 38] is that $\mathbb{E}(R^{(\mathrm{al})}_{n,n}) = O(\sqrt{n})$, but the complexity of the random variable does not allow for direct calculations. In this article we obtain an asymptotic lower bound for the optimal score when a is fixed, as well as upper bounds for the number of optimality regions when the rectangle is of dimensions $m_n \times n$. With random words of this size the biological applications are unrealistic but the results offer some insight from a theoretical perspective. Moreover, we prove that $O(\sqrt{n})$ for the expectation is not the correct order in this case, at least for $a < 3/4$.

Optimality regions can be studied in the independent model as well, and in fact we can obtain stronger results, again when the rectangle is of dimensions $m_n \times n$.

1.5 Outline

The paper is organised as follows: in Section 2 we state our main results. Section 3 contains preliminary results that do not depend on the specific choice of environment and therefore hold for both the alignment and the independent model. The results concerning the independent model are proved in Section 4 whereas in Section 5 we prove our results about the alignment model.

1.6 Notation

We briefly collect the pieces of notation discussed so far and list the most common notation used in the paper. Letters T, G and L all denote last passage times: T is for passage times under potential (1.2), G is the passage time for the independent model and L its counterpart for the alignment model. The letter R is reserved for the number of optimality regions, and we distinguish the regions in each of the two models by $R^{(\mathrm{ind})}_{m,n}$ the regions in the independent model, and by $R^{(\mathrm{al})}_{m,n}$ the regions in the alignment model. We omit the superscripts when results hold for both models (see for example Section 3).

Throughout, p is a parameter in the interval $(0, 1)$ and $q = 1-p$. $\mathcal{A}$ is the alphabet in the alignment model and $|\mathcal{A}|$ is its size.

2 Results

In this section we have our main results, first for the independent model and then the softer ones for the alignment model.

2.1 Independent Model

See Section 4 for a proof of Theorems 2.1, 2.3, 2.4 and Corollary 2.2.

We consider the last passage time $G^{(0)}_{m_n,n}$ with $m_n = n/p - xn^a$ for suitably chosen x. When the exponent a is small we obtain tightness without rescaling, for any choice of x:

Theorem 2.1 *Let $x \in \mathbb{R}$ and $a \in \left(0, \frac{1}{2}\right]$. The sequence $\left(n - G^{(0)}_{n/p-xn^a,n}\right)_{n\in\mathbb{N}}$ is tight and*

$$\overline{\lim_{n\to\infty}} \mathbb{P}\left\{n - G^{(0)}_{n/p-xn^a,n} \geq k\right\} \leq \begin{cases} 2^{-k}, & a < 1/2 \\ \left(\Phi\left(xpq^{-\frac{1}{2}}\right)\right)^k, & a = 1/2. \end{cases} \tag{2.1}$$

where Φ is the cumulative distribution function of the standard Gaussian distribution.

We will see in (3.12) that $R^{(\mathrm{ind})}_{m,n} < n - G^{(0)}_{m,n}$. As a corollary we obtain an asymptotic bound on the expected number of optimality regions. A Monte Carlo Simulation for the expected number of regions can be seen in Fig. 2.

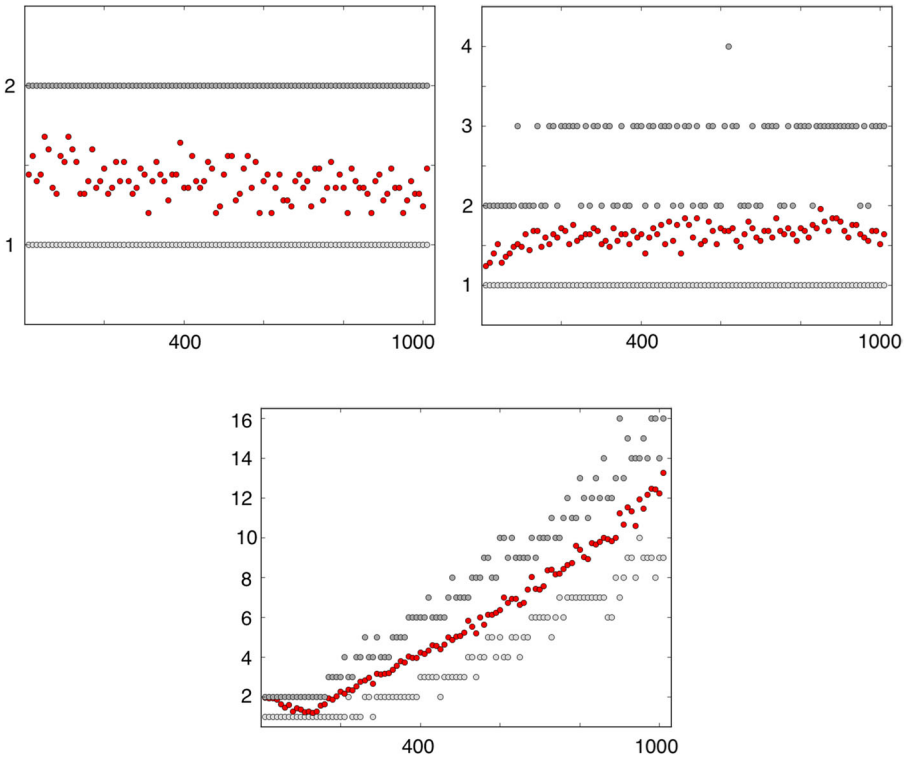

Fig. 2 Monte Carlo simulations for the empirical maximum, minimum and expected number of regions for up to $n = 1000$ in the independent model for $a = 0.8$ with varying $p = 0.05, 0.5, 0.8$ from left to right. For each n, 25 independent environments were sampled. n grows in increments of size 10

Corollary 2.2 *Let* $a \in \left(0, \frac{1}{2}\right]$. *Then*

$$\overline{\lim_{n\to\infty}} \mathbb{E}\left[R^{(\mathrm{ind})}_{n/p-xn^a,n}\right] \leq \begin{cases} 2, & a < 1/2 \\ \left(1-\Phi\left(xpq^{-\frac{1}{2}}\right)\right)^{-1}, & a = 1/2. \end{cases} \tag{2.2}$$

For $a > 1/2$ we state a bound on the number $R^{(\mathrm{ind})}_{m,n}$ of optimality regions. The optimal results and the relevant scaling of m in terms of n differ according to the value of a.

Theorem 2.3 *Let* $a \in (0, 1)$.

(1) *If* $a \in \left(0, \frac{1}{2}\right]$,

$$\overline{\lim_{n\to\infty}} \mathbb{P}\left\{R^{(\mathrm{ind})}_{n/p-xn^a,n} \geq k\right\} \leq \begin{cases} 2^{-k}, & a < 1/2 \\ \left(\Phi\left(xpq^{-\frac{1}{2}}\right)\right)^{k}, & a = 1/2. \end{cases} \tag{2.3}$$

(2) *If* $a \in \left(\frac{1}{2}, \frac{3}{4}\right]$ *there exists a constant* $C_1 = C_1(x, p)$ *so that*

$$\lim_{n\to\infty} \mathbb{P}\left\{R^{(\mathrm{ind})}_{n/p-xn^a,n} > C_1 n^{2a-1}\right\} = 0. \tag{2.4}$$

(3) *If* $a \in \left(\frac{3}{4}, 1\right)$ *there exists a constant* $C_2 = C(x, p)$ *so that,*

$$\lim_{n\to\infty} \mathbb{P}\left\{R^{(\mathrm{ind})}_{n/p-xn^a,n} > C_2 n^{2a/3}\right\} = 0. \tag{2.5}$$

In the theorem above, (2.4) holds also when $a > 3/4$, however the bound $n^{2a/3}$ is sharper.

Finally when $a \in \left(\frac{1}{2}, \frac{5}{7}\right]$ we obtain Tracy-Widom fluctuations. It is worth noting that we do not take the standard approach of scaling by the variance. Instead, we change the size of the rectangle, by subtracting a term of size $n^{\frac{2-a}{3}}$ from the width (Fig. 3).

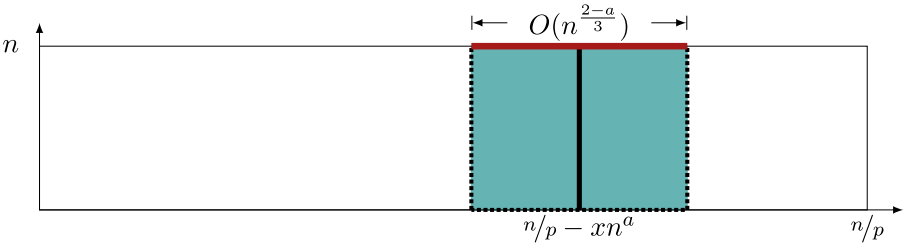

Fig. 3 Tracy-Widom fluctuations to the last passage time of the independent model depend on position of the endpoint in the thickset red line. When $a \in \left(\frac{1}{2}, \frac{2}{3}\right)$ the Tracy-Widom reveals itself just by centering according to the first and second order macroscopic approximation of the LLN for G. However when $a \in (\frac{3}{2}, \frac{5}{7})$, a third order approximation to the law of large numbers, cn^{3a-2}, is necessary for the Tracy-Widom fluctuations

Theorem 2.4 *For $s \in \mathbb{R}$ define $x = \frac{2}{\sqrt{p}}\left(\frac{q}{p}\right)^a$ and $y(s) = s\frac{\sqrt{p}}{q}\left(\frac{p}{q}\right)^{\frac{1+a}{3}}$. Then*

(1) *For $1/2 < a < 2/3$,*

$$\lim_{n\to\infty} \mathbb{P}\left\{ G^{(0)}_{\frac{n}{p}-xn^a-y(s)n^{\frac{2-a}{3}},n} \le n - \left(\frac{qn}{p}\right)^{2a-1} \right\} = F_{TW}(s). \tag{2.6}$$

(2) *For $2/3 \le a \le 5/7$,*

$$\lim_{n\to\infty} \mathbb{P}\left\{ G^{(0)}_{\frac{n}{p}-xn^a-y(s)n^{\frac{2-a}{3}},n} \le n - \left(\frac{qn}{p}\right)^{2a-1} + ax\left(\frac{q}{p}\right)^{2a-2} n^{3a-2} \right\} = F_{TW}(s). \tag{2.7}$$

Remark 2.5 The case $a \ge \frac{5}{7}$ corresponds to an exponent $\gamma = 2a - 1 \ge \frac{3}{7}$ in (1.5) (see [7]) and the result cannot be extended further with these techniques. In Section 3.1 of [7] the authors explain why their result should extend at least up to exponent $\gamma = \frac{3}{4}$. The independent Bernoulli model here, while equivalent to the edge of the corner growth model may be a bit more sensitive to these cut-offs and indeed $\gamma = \frac{3}{7}$ seems to be critical and manifests itself in the proof.

From the two cases of Theorem 2.4 we see that we need to amend the right-hand side of the event in (2.6) by a term $O(n^{3a-2})$, in order to get the non-trivial result in (2.7). This gives a new cut-off $a = \frac{2}{3}$ $\left(\text{or } \gamma = \frac{1}{3}\right)$. The term is there for case 2 as well, but when $a \le \frac{2}{3}$ the term is bounded and plays no role, while it must be dealt with, for higher a.

Second, from the proof of Theorem 2.4, the exponent $a = \frac{5}{7}$ $\left(\gamma = \frac{3}{7}\right)$ seems to be critical, since it is necessary to have $2a - 1 < \frac{2-a}{3}$ to balance the various orders of magnitude that appear. Assuming that the scaling in (1.5) remains the same for $\gamma \in \left(\frac{3}{7}, \frac{3}{4}\right)$, this change implies a corresponding correction term of size $O(n^\gamma)$ at the numerator of (1.5).

2.2 Alignment Model

Throughout we fix a finite alphabet $\mathcal{A}$ with $|\mathcal{A}| \ge 2$, from which the letters of words η^{x} and η^{y} are chosen uniformly at random, independently of each other and let $a \in (0, 1)$ and $\alpha, \beta \ge 0$. The proofs of Theorems 2.6, 2.7 and 2.8 can be found in Section 5.

Define

$$g^{(a)}(n) = \begin{cases} \sqrt{n \log n}, & a \le 1/2, \\ n^a, & a > 1/2. \end{cases}$$

Theorem 2.6 *Let $x > 0$ and $\beta \ge 0$. For $\mathbb{P}$-a.e. ω we have the upper bound*

$$\overline{\lim_{n\to\infty}} \frac{n(1+\beta-\beta|\mathcal{A}|) - L^{(\beta)}_{\lfloor n|\mathcal{A}|-xn^a\rfloor,n}}{g^{(a)}(n)} \le \begin{cases} \frac{\sqrt{2}}{|\mathcal{A}|-1} - \frac{1}{|\mathcal{A}|}, & a \le 1/2, \\ \frac{1}{|\mathcal{A}|(|\mathcal{A}|-1)} - \beta x, & a > 1/2. \end{cases}$$

and the lower bound

$$\varliminf_{n\to\infty} \frac{n(1+\beta-\beta|\mathcal{A}|) - L^{(\beta)}_{\lfloor n|\mathcal{A}|-xn^a\rfloor,n}}{g^{(a)}(n)} \geq \begin{cases} 0, & a \leq 1/2, \\ -\beta x, & a > 1/2. \end{cases}$$

Finally, we turn to the number of optimality regions for the alignment model. The first result gives an upper bound on the asymptotic growth of the number regions:

Theorem 2.7 *Let $x > 0$. There exists a constant $C_1(|\mathcal{A}|, x)$ that only depend on x and $|\mathcal{A}|$ so that*

$$\varlimsup_{n\to\infty} \frac{R^{(\mathrm{al})}_{\lfloor|\mathcal{A}|n-xn^a\rfloor}n}{(g^{(a)}(n))^{\frac{2}{3}}} \leq C_1(x, |\mathcal{A}|), \quad \mathbb{P}-a.s. \tag{2.8}$$

The constant tends to 0 as the alphabet size tends to ∞.

We also have a bound of the same order for the expected number of optimality regions.

Theorem 2.8 *There exists a constant $C_2(x, |\mathcal{A}|)$, i so that*

$$\varlimsup_{n\to\infty} \frac{\mathbb{E}\left[R^{(\mathrm{al})}_{|\mathcal{A}|n-xn^a}n\right]}{(g^{(a)}(n))^{\frac{2}{3}}} \leq C_2(x, |\mathcal{A}|). \tag{2.9}$$

The constant tends to 0 as the alphabet size tends to ∞.

Remark 2.9 These results are also valid for the independent model. Given the stronger bounds for the independent model, we do not expect (2.9) to be sharp, particularly for small values of the exponent a, and this is supported by Monte Carlo simulations. For example these suggest that for $a \leq 1/2$ the number of expected regions is bounded (see Fig. 4). This is also the case for the independent model as we see in Theorem 2.2. For $a > 1/2$, the simulations in Fig. 5 show that the expected number of regions is growing for small alphabet sizes, but again the exponent of growth is smaller than $2a/3$ and it seems to depend on the alphabet size.

3 Model Independent Results for Optimality Regions and Maximal Paths

In this section we present preliminary results about the two models that do not depend on the correlation structure of the weights. We therefore write $R_{m,n}$ to mean either $R^{(\mathrm{al})}_{m,n}$ or $R^{(\mathrm{ind})}_{m,n}$. We also introduce the vocabulary usually used in the sequence alignment literature.

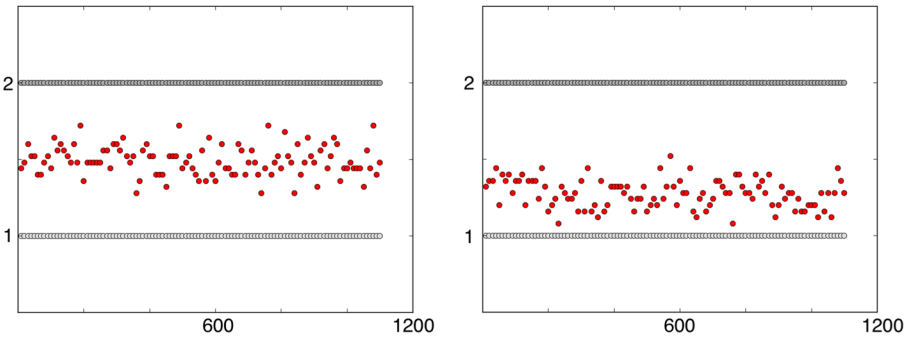

Fig. 4 Monte Carlo simulations for the empirical maximum, minimum and expected number of regions for up to $n = 1000$ in the alignment model for small values of a. For each n, 25 independent pairs of strings were uniformly chosen. n grows in increments of size 10
(Left) $|\mathcal{A}| = 20, a = 1/3, x = 1$. (Right) $|\mathcal{A}| = 2, a = 1/2, x = 1$. The simulations suggest the expected number of regions is bounded, and in agreement with the theoretical bound obtained for the independent model

Let $\pi = \{u_0, \ldots, u_M\} \in \Pi_{0,(m,n)}$ denote an admissible path and recall that the increments $z_k = u_k - u_{k-1} \in \mathcal{R} = \{e_1, e_2, e_1 + e_2\}$. Thus for each increment there are three possibilities:

(1) $z_k = e_1 + e_2$ with $\omega_{u_k} = 0$, called a *mismatch*,
(2) $z_k \in \{e_1, e_2\}$, called a *gap*,
(3) $z_k = e_1 + e_2$ with $\omega_{u_k} = 1$, called a *match*.

Let $x = x(\pi)$ be the number of mismatches, $y = y(\pi)$ the number of gaps and $z = z(\pi)$ the number of matches of π. We also denote this triplet by $\mathbf{s}(\pi) = (x(\pi), y(\pi), z(\pi))$.

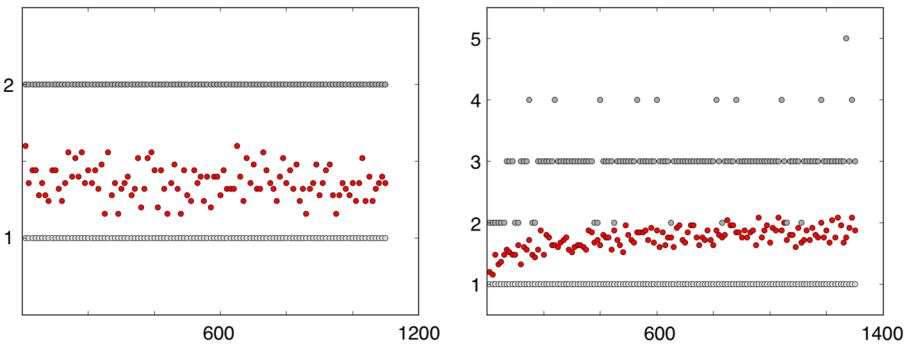

Fig. 5 Monte Carlo simulations for the empirical maximum, minimum and expected number of regions in the alignment model when a is close to 1. For each n, 25 independent pairs of strings were uniformly chosen. n grows in increments of size 10. (Left) $|\mathcal{A}| = 20, a = 0.8, x = 1$. (Right) $|\mathcal{A}| = 2, a = 0.8, x = 1$. The simulations suggest that the expected number of regions is bounded for large alphabet sizes, but for small size alphabets we see growth

Fix parameters $\alpha, \beta \geq 0$. Under potential $V_{\alpha,\beta}$ the score of the path π is then given by

$$w_{\alpha,\beta}(\pi) = z - \alpha x - \beta y. \tag{3.1}$$

Since any diagonal step is equivalent to an e_1 step followed by a e_2 step or vice versa, we have

$$m + n = 2x(\pi) + 2z(\pi) + y(\pi) \qquad \text{for all } \pi \in \Pi_{0,(m,n)}. \tag{3.2}$$

The last passage time $G^{(\alpha,\beta)}_{m,n}$ (or $L^{(\alpha,\beta)}_{m,n}$, depending on the environment) under potential defined in (1.3) can now be rewritten as

$$G^{(\alpha,\beta)}_{m,n} = \max_{\pi \in \Pi_{0,(m,n)}} \{w_{\alpha,\beta}(\pi)\}.$$

Our focus will be on the *minimal-gap maximisers (MGM)*: paths whose score attains the last passage time with the smallest possible number of gaps. Since any two MGM paths have the same number of gaps and the same score it follows from (3.2) that

Lemma 3.1 *All MGM paths have the same number of gaps, matches and mismatches.*

We denote the set of MGM paths by $\Gamma^{(\alpha,\beta)}_{0,(m,n)}$. When $\alpha = 0$ we write $\Gamma^{(\beta)}_{0,(m,n)}$.

Definition 3.2 Two points (α_1, β_1) and (α_2, β_2) belong in different *optimality regions of the parameter space* for a fixed terminal point (m, n) if and only if $\Gamma^{(\alpha_1,\beta_1)}_{0,(m,n)} \cap \Gamma^{(\alpha_2,\beta_2)}_{0,(m,n)} = \varnothing$.

For future reference we record the following observations:

(1) For fixed $\alpha \geq 0$ and any $\beta_1 \leq \beta_2$ we have

$$w_{\alpha,\beta_1}(\pi) \geq w_{\alpha,\beta_2}(\pi) \tag{3.3}$$

and therefore this inequality also holds for the passage times:

$$G^{(\alpha,\beta_1)}_{m,n} \geq G^{(\alpha,\beta_2)}_{m,n} \qquad \text{and} \qquad L^{(\alpha,\beta_1)}_{m,n} \geq L^{(\alpha,\beta_2)}_{m,n} \tag{3.4}$$

(2) For $\alpha = -1$ and $\beta = -1/2$, the weight of any path $\pi \in \Pi_{0,(m,n)}$ is given by

$$w_{-1,-1/2}(\pi) = \frac{m+n}{2} \tag{3.5}$$

Lemma 3.3 *All optimality regions in the (α, β)-positive quadrant are semi-infinite cones bounded by the coordinate axes and lines of the form $\beta = c + \alpha\left(c + \frac{1}{2}\right)$.*

This result was first proved in [19]; we give a simplified proof here:

Proof Pick any $(\alpha, \beta) \in \mathbb{R}^2_+$ and let $(0, \beta')$ be the point of intersection of the linear segment connecting (α, β) and $(-1, -1/2)$ with the y-axis, i.e.

$$\beta = (\alpha + 1)\beta' + \frac{\alpha}{2}. \tag{3.6}$$

We will show that the optimal paths associated with $(0, \beta')$ are the same as those associated to (α, β). Consider any $\pi \in \Pi_{0,(m,n)}$ with $\mathbf{s}(\pi) = (x, y, z)$. Then

$$\begin{aligned} w_{\alpha,\beta}(\pi) &= z - x\alpha - y\beta = z - x\alpha - y\beta + y\beta' - y\beta' = w_{0,\beta'}(\pi) - x\alpha - (\beta - \beta')y \\ &= w_{0,\beta'}(\pi) - x\alpha - \Big((\alpha+1)\beta' + \frac{\alpha}{2} - \beta'\Big)y, \quad \text{by (3.6)}, \\ &= w_{0,\beta'}(\pi) - \alpha\left(\frac{m+n}{2} - w_{0,\beta'}(\pi)\right), \quad \text{by (3.2)}, \\ &= (1+\alpha)w_{0,\beta'}(\pi) - \alpha\frac{m+n}{2}. \end{aligned} \tag{3.7}$$

So the weight of any path with parameters (α, β) is an affine function of the weight with parameters $(0, \beta')$ and the two parameters must belong to the same optimality region. □

Under a fixed environment ω, we define the *critical penalties*

$$0 < \beta_1 < \cdots < \beta_{R_{m,n}} < \infty \tag{3.8}$$

to be the the gap penalties for $\alpha = 0$ at which the optimality region changes. We will also write β_∞ for the last threshold $\beta_{R_{m,n}}$.

Lemma 3.4 (Critical penalties) *For each $k \le R_{m,n}$ let $\pi^{(\beta_k)} \in \Gamma^{(\beta_k)}_{0,(m,n)}$, with $\mathbf{s}(\pi^{(\beta_k)}) = (x_{\beta_k}, y_{\beta_k}, z_{\beta_k})$. Then*

$$\beta_{k+1} = \frac{z_{\beta_k} - z_{\beta_{k+1}}}{y_{\beta_k} - y_{\beta_{k+1}}}. \tag{3.9}$$

Proof Continuity of the optimal score in the parameter β implies that at β_{k+1} the weights will be the same whether β_{k+1} is approached by above (considering scores of paths in $\Gamma^{(\beta_{k+1})}_{0,(m,n)}$) or from below (scores of paths in $\Gamma^{(\beta_k)}_{0,(m,n)}$). Therefore

$$z_{\beta_k} - \beta_{k+1} y_{\beta_k} = z_{\beta_{k+1}} - \beta_{k+1} y_{\beta_{k+1}}$$

which yields the conclusion. □

Upper bounds for the maximal value of $R_{m,n}$ can be found in [11]. For the LCS model these are sharp when the alphabet size grows to infinity. The results and arguments in [11] can be extended to give the upper bound

$$R_{\lfloor ns \rfloor + o(n), \lfloor nt \rfloor + o(n)} \le Cn^{2/3}, \tag{3.10}$$

that holds in any fixed realization of the environment, any $(s, t) \in \mathbb{R}^2_+$ and n large enough. They also proved that environments that actually generate so many regions exist, at least when the alphabet size was infinite. This was later verified also for finite alphabets in [45].

Lemma 3.5 *For $\beta_0 = 0$ and each critical β_k in (3.8), choose an MGM path $\pi_i \in \Gamma^{(\beta_i)}_{0,(m,n)}$ with $\mathbf{s}(\pi_i) = (x_i, y_i, z_i)$ for $0 \le i \le R_{m,n}$. Then*

$$R_{m,n} \le \min\left\{z_0 - z_{R_{m,n}}, \frac{x_{R_{m,n}} - x_0}{2}, \frac{y_0 - y_{R_{m,n}}}{2}, n \wedge m - z_0\right\}. \tag{3.11}$$

Proof Since the paths π_i correspond to different penalties β_i, they must differ in the number of diagonal steps and the number of gaps. Since a diagonal step is equivalent to two gaps, we have $y_i - y_{i+1} \ge 2$. Furthermore it must be the case that $z_i - z_{i+1} \ge 1$; otherwise π_i would violate the MGM condition. Equation (3.2) and the last two inequalities give $x_{i+1} - x_i \ge 2$. Adding each inequality over i gives the first three terms in the minimum of (3.11). For the last term note that $y_{R_{m,n}} = n \vee m - m \wedge n$. Since $x_0 \ge 0$ (3.2) yields $2(n \wedge m - z_0) \ge y_0 - y_{R_{m,n}}$. □

Remark 3.6 Notice that the last bound in (3.11) can be written as

$$R^{(\mathrm{al})}_{m,n} \le n - L^{(0)}_{m,n} \qquad \text{and} \qquad R^{(\mathrm{ind})}_{m,n} \le n - G^{(0)}_{m,n}. \tag{3.12}$$

Finally, we present a lemma that gives a useful bound on the number of regions if a bit more information is available.

Lemma 3.7 *Let $m = m(n)$ so that $m(n) \to \infty$ as $n \to \infty$. Let $g(n)$ be a deterministic function so that $\lim_{n\to\infty} g(n) = \infty$. Then, there exists an $N > 0$ and a non-random constant C_0 so that for all $n > N$ we have the inclusion of events*

$$A_n = \{z_0 - z_R + y_0 - y_R \le g(n)\} \subseteq \{R_{m,n} \le C_0 (g(n))^{\frac{2}{3}}\}. \tag{3.13}$$

In particular,

(1) *If $\mathbb{P}\{A_n^c \ i.o\} = 0$, then the number of optimality regions $R_{m,n}$ satisfies*

$$\overline{\lim_{n\to\infty}} \frac{R_{m,n}}{g(n)^{\frac{2}{3}}} \le C_0, \quad \mathbb{P} - a.s. \tag{3.14}$$

(2) *If $\mathbb{P}\{A_n\} \to 1$, then the number of optimality regions $R_{m,n}$ satisfies*

$$\lim_{n\to\infty} \mathbb{P}\left\{\frac{R_{m,n}}{g(n)^{\frac{2}{3}}} \le C_0\right\} = 1. \tag{3.15}$$

Proof Statements (3.14), (3.15) are immediate corollaries of (3.13) which we now show. Fix an environment $\omega \in A_n$. Then we have that

$$z_0 - z_R + y_0 - y_R = \sum_{i=0}^{R_{m,n}-1} \{(z_{\beta_{i+1}} - z_{\beta_i}) + (y_{\beta_{i+1}} - y_{\beta_i})\} \le g(n).$$

The sum above has as terms the numerators and denominators of the critical penalties (see Lemma 3.4).

Each critical penalty is a distinct rational number and it corresponds to a change of optimality region. The bound $g(n)$ is independent of the environment, so we

can obtain an upper bound on the number of regions that is independent of the environment, if we maximize the number of terms that appear in the sum.

Since the terms in the sum are integers, the maximal number of terms is the maximal number of integers k that can be added so that the bound $g(n)$ is not exceeded. Those integers k need not be distinct but they need to able to be written as a sum of integers a, b, $k = a + b$ so that a/b are different. This is because the ratio a/b corresponds to critical penalties and those are distinct. Take each successive integer k and compute the number of irreducible fractions a/b so that $a + b = k$.

The number of irreducible fractions satisfying this is $\varphi(k)$, where φ is Euler's totient function [3]. The number of distinct values k that can be used is $M_{\max}$, which must satisfy

$$\sum_{k=1}^{M_{\max}} k\varphi(k) \leq g(n) < \sum_{k=1}^{M_{\max}+1} k\varphi(k).$$

These inequalities imply that $M_{\max}$ will be bounded above, up to a lower order term, by $cg(n)^{1/3}$. This follows by the asymptotics of φ for large arguments, and we direct the reader to the proof of Theorem 5 in [11] for the details. The bound on $M_{\max}$ is true for all $n > N_1$ large enough. Then an upper bound for the number of admissible pairs (a, b) (and therefore for the maximal number of regions) is

$$\sum_{k=1}^{M_{\max}} \phi(k) \leq c_1 M_{\max}^2 \leq Cg(n)^{2/3}.$$

This last estimate is again the result of an analytic number theory formula (see [3]) which also works for $n > N_2$ large enough. So both deterministic bounds hold for all $n > N = N_1 \vee N_2$. □

The difficulty with the alignment model is the correlated environment. Therefore, the soft techniques below try to avoid precisely this issue. The same techniques work for the independent model and give identical bounds, but the exact solvability of that model often allows sharper results.

Our strategy is to construct a path with a score that is near-optimal under any penalty β and which attempts to minimize as much as possible the number of vertical steps. This will be important for the lower bound for the passage time under penalty β_R, where we know that the optimal path takes no vertical steps. We present the construction and results for alignment model, but re-emphasize that they hold for both.

3.1 Construction of the Path

Fix an environment ω on $\mathbb{N}^2$, defined by two infinite words η^{x}, η^{y}, where each letter is chosen uniformly at random. $\omega_{i,j}$ is defined according to (1.6).

Consider the following strategy (S) to create a path π_S:

(1) For some appropriate constants c_1 and c_2 (to be determined later), move with $e_1 + e_2$ steps from 0 up to a fixed point

$$u_n(a) = \begin{cases} \left(\left\lfloor \sqrt{c_1 n \log n} \right\rfloor, \left\lfloor \sqrt{c_1 n \log n} \right\rfloor \right), & \text{if} a \leq \frac{1}{2}, \\ \left(\left\lfloor \frac{1}{|\mathcal{A}|-1} x n^a \right\rfloor + \left\lfloor \sqrt{c_2 n \log n} \right\rfloor, \left\lfloor \frac{1}{|\mathcal{A}|-1} x n^a \right\rfloor + \left\lfloor \sqrt{c_2 n \log n} \right\rfloor \right), & \text{if} a > \frac{1}{2}. \end{cases} \tag{3.16}$$

(2) Now, construct the path as follows, depending on the current position $u_n(a)$:

(a) If the path is on site (i, j) with $j < n$ and $\omega_{i+1,j+1} = 1$ then move diagonally with an $e_1 + e_2$ step, and now the path is on site $(i + 1, j + 1)$.
(b) If the path is on site (i, j) with $i < \lfloor |\mathcal{A}| n - x n^a \rfloor$ and $\omega_{i+1,j+1} = 0$ then move horizontally with an e_1 step, and now the path is on site $(i + 1, j)$.
(c) If $j = n$ or $i = \lfloor |\mathcal{A}| n - x n^a \rfloor$, move to $(\lfloor |\mathcal{A}| n - x n^a \rfloor, n)$.

From this description it is not clear whether we can enforce the condition that no vertical steps will be taken by π_S. However, this will happen for eventually all n, by choosing constants c_1, c_2 appropriately. Consider an infinite path $\bar{\pi}_S$ that moves according to strategy (S) but without the restrictions $i < \lfloor |\mathcal{A}| n - x n^a \rfloor$ for (3)-(b) and without step (3)-(c).

Let Y_j be the random variables that give the amount of horizontal steps path $\bar{\pi}_S$ takes at level $y = j + u_n(a) \cdot e_2$,

$$Y_j = |\{i \in \mathbb{N} : (i, j + u_n(a) \cdot e_2) \in \bar{\pi}_S\}|. \tag{3.17}$$

Because $\bar{\pi}_S$ does not have a target endpoint, we have

$$Y_j \sim \text{Geom}\left(\frac{1}{|\mathcal{A}|}\right), \quad \mathbb{P}\{Y_j = \ell\} = \frac{1}{|\mathcal{A}|}\left(1 - \frac{1}{|\mathcal{A}|}\right)^{\ell-1}. \tag{3.18}$$

By construction, the Y_j are i.i.d. with mean $|\mathcal{A}|$.

Path $\bar{\pi}_S$ coincides with π_S up until the point that $\bar{\pi}_S$ hits either the north or east boundary of the rectangle $[0, \lfloor |\mathcal{A}| n - x n^a \rfloor] \times [0, n]$. When $\bar{\pi}_S$ touches the north boundary first, we can conclude that π_S has no vertical steps up to that point. We will estimate precisely this probability, using the following moderate deviations lemma [9].

Lemma 3.8 *Let $(X_N)_{N\in\mathbb{N}}$ an i.i.d. sequence of random variables with exponential moments. If $N\lambda_N^2 \to \infty$ and $N\lambda_N^3 \to 0$ then*

$$\mathbb{P}\left\{ \left| \frac{1}{N} \sum_{i=1}^{N} X_i - \mathbb{E}(X_1) \right| > \lambda_N \right\} \sim \frac{2}{\sqrt{2\pi N \lambda_N^2}} e^{-N\lambda_N^2/2}. \tag{3.19}$$

From the equality of events

$$\{\bar{\pi}_S \text{ exits from the north boundary}\} = \left\{ \sum_{j=1}^{n-u_n(a)\cdot e_2} Y_j \le \lfloor |\mathcal{A}|n - xn^a \rfloor - u_n(a)\cdot e_1 \right\}, \tag{3.20}$$

we estimate for $a \le 1/2$ and for n sufficiently large for the asymptotics in (3.19) to be accurate,

$$\begin{aligned}
&\mathbb{P}\left\{ \sum_{j=1}^{n-u_n(a)\cdot e_2} Y_j \le \lfloor |\mathcal{A}|n - xn^a \rfloor - \left\lfloor \sqrt{c_1 n \log n} \right\rfloor \right\} \\
&\ge \mathbb{P}\left\{ \sum_{j=1}^{n-\left\lfloor \sqrt{c_1 n \log n} \right\rfloor} (Y_j - |\mathcal{A}|) \le -xn^a + (|\mathcal{A}|-1)\sqrt{c_1 n \log n} - 3 \right\} \\
&\ge 1 - c_0 \frac{1}{\sqrt{\log n}} n^{-c_1(|\mathcal{A}|-1)^2/4}.
\end{aligned} \tag{3.21}$$

For the last inequality, we used Lemma 3.8 for

$$N = n - \left\lfloor \sqrt{c_1 n \log n} \right\rfloor \text{ and} \lambda_N = (|\mathcal{A}|-1)\sqrt{c_1}\sqrt{\frac{\log n}{n}} + O(n^{\alpha-1}).$$

The constant c_0 only depends on $|\mathcal{A}|$ which is assumed to be strictly larger than 1. Choose $c_1 > \frac{2}{(|\mathcal{A}|-1)^2}$ so that the probabilities of the event $\{\bar{\pi}_S \text{ exits from the east boundary}\}$ are summable in n. Then by the Borel-Cantelli lemma, we can find an $M = M(\omega)$ so that for all $n > M$ path $\bar{\pi}_S$ hits the north boundary first.

The situation for $a > 1/2$ is similar. Starting from (3.21), we have

$$\begin{aligned}
&\mathbb{P}\left\{ \sum_{j=1}^{n-u_n(a)\cdot e_2} Y_j \le \lfloor |\mathcal{A}|n - xn^a \rfloor - \left\lfloor \frac{1}{|\mathcal{A}|-1} xn^a \right\rfloor - \left\lfloor \sqrt{c_2 n \log n} \right\rfloor \right\} \\
&\ge \mathbb{P}\left\{ \sum_{j=1}^{n-\left\lfloor \frac{1}{|\mathcal{A}|-1} xn^a \right\rfloor - \left\lfloor \sqrt{c_2 n \log n} \right\rfloor} (Y_j - |\mathcal{A}|) \le (|\mathcal{A}|-1)\sqrt{c_2 n \log n} - 3 \right\}.
\end{aligned} \tag{3.22}$$

Then the proof goes as for the previous case, and again it suffices that $c_2 > \frac{2}{(|\mathcal{A}|-1)^2}$.

From the definition of π_S and the above discussion, we have shown the following:

Lemma 3.9 *For $\mathbb{P}$- a.e. ω there exists $M = M(\omega)$ so that for all $n > M(\omega)$, path π_S exits from the north boundary of the rectangle $[0, |\mathcal{A}|n - xn^a] \times [0, n]$. In that case,*

(1) *it has no vertical gaps until the point of exit,*
(2) *the number of horizontal gaps it has is $(|\mathcal{A}| - 1)n - xn^a$ (the minimal possible), and*
(3) *it collects $n - u_n(a) \cdot e_2 + \sum_{k=1}^{u_n(a)\cdot e_2} \omega_{k,k}$ positive weight.*

Since π_S has the smallest number of gaps possible, it can be optimal under any penalty β.

4 The Independent Model

In this section we prove results about the independent model. We begin with a coupling between the longest common subsequence in the independent model, with the corner growth model in an i.i.d. Geom$(1 - p)$ environment. This is achieved via the following identity. Recall that $T_{m,n}$ denotes the last passage time in an $m \times n$ rectangle, with admissible e_1 or e_2 steps only, under potential (1.2).

$$\mathbb{P}\left\{G^{(0)}_{m,n} \leq m - N\right\} = \mathbb{P}\left\{T_{n-m+N,N} \leq n + N - 1\right\}. \tag{4.1}$$

The result follows from the arguments in [13], and we briefly present the main idea.

The *discrete totally asymmetric simple exclusion process* (DTASEP) with backward updating is an interacting particle system of left-finite particle configuration on the integer lattice, i.e. such that sites to the left of some threshold are empty (see Fig. 6). Label the particles from left to right and denote the position of the j^{th} particle

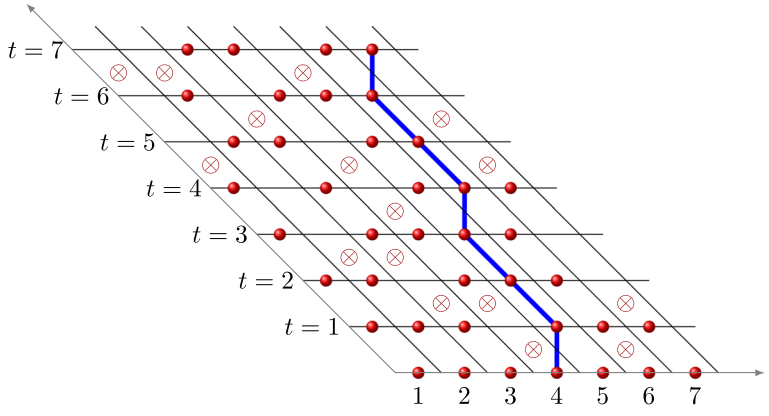

Fig. 6 Space-time realisation of DTASEP (Graphical construction). Particles move to the left, according to exclusion rules (1) and (2). Symbols $\otimes$ denote Bernoulli(p) weights 1, and particle underneath an $\otimes$ symbol cannot jump during that time, i.e. particles jump with probability $1-p = q$ as long as the exclusion rule is not violated. The trajectory of particle 4 is highlighted for reference

at time $\ell \in \mathbb{N}$ by $\eta_j(\ell)$. At every discrete time step $\ell \in \mathbb{N}$ each particle independently attempts to jump one step to the left with probability $q = 1 - p$. Particle i performs the jump if either

(1) the target site was unoccupied by particle $i - 1$ at time $\ell - 1$ or,
(2) the target site was occupied by particle $i - 1$, but it also performs a jump at time ℓ.

In words, particles are forbidden to jump to occupied sites and we update from left to right. Start DTASEP with the *step initial condition* $\eta_i(0) = i$ so that initially the i-th particle is at position i. Let $\tau_{i,j}$ be the time it takes particle j to jump i times:

$$\tau_{i,j} = \inf\{\ell \geq 0 : \eta_j(\ell) \leq j - i\}.$$

Then the following recursive equation holds

$$\tau_{i,j} = \tau_{i,j-1} \vee (\tau_{i-1,j} + 1) + \tilde{\zeta}_{i,j}.$$

where the $\tilde{\zeta}_{i,j}$ are independent Geometric variables with parameter $q = 1 - p$, supported on $\mathbb{N}_0$.

By setting $\zeta_{i,j} = \tilde{\zeta}_{i,j} + 1 \sim \text{Geom}(1 - p) \in \{1, 2, \ldots\}$, the $\tau_{i,j}$ can be coupled with the last passage time in the corner growth model (cf. [13], Lemma 5.1), giving the equality in distribution

$$\tau_{i,j} \overset{\text{(d)}}{=} T_{i,j} - j + 1. \tag{4.2}$$

We embed DTASEP in the two-dimensional lattice $\mathbb{Z} \times \mathbb{N}_+$, using its graphical construction as follows: Let $\left\{b_{k,\ell} \colon (k, \ell) \in \mathbb{Z} \times \mathbb{N}_+\right\}$ be a field of i.i.d. Bernoulli(q) random variables and assign to each site (k, ℓ) the random weight $b_{k,\ell}$. Particles are placed initially on $\mathbb{N}_+ \times \{0\}$, with particle i at coordinate $(\eta_i(0), 0)$. The Bernoulli marked sites signify which particles will attempt to jump in the DTASEP process.

After the spatial locations in the DTASEP at time $\ell = 1$ are determined, the particles in the graphical construction are at positions $(\eta_i(1), 1)$. We iterate this procedure for all times $\ell \in \mathbb{N}$.

Then, the environments between graphical DTASEP and BLIP may be coupled via

$$1 - \omega_{k,\ell} = b_{k+\ell,\ell}.$$

In [13] the following combinatorial identity was proved:

$$G^{(0)}_{m,n} = m - \max\{k : (m - n) \vee 1 \leq k \leq m,\ \tau_{k+n-m,k} \leq n\}. \tag{4.3}$$

Set $k^* = \max\{k \leq m : k \geq (m - n) \vee 1,\ \tau_{k+n-m,k} \leq n\} \vee 0$. Then

$$\{G^{(0)}_{m.n} \leq m - N\} = \{N \leq k^*\} = \{\tau_{N+n-m,N} \leq n\}, \tag{4.4}$$

where the last equality comes form the fact that $\tau_{N+n-m,N}$ is an increasing random variable in N For a clear pictorial explanation about the coupling, look at Fig. 7. Finally compute

$$\begin{aligned}
\mathbb{P}\left\{G^{(0)}_{m,n} \leq m - N\right\} &= \mathbb{P}\left\{N \leq \max\{k : (m - n) \vee 1 \leq k \leq m,\ \tau_{k+n-m,k} \leq n\}\right\}, \quad \text{by (4.3)} \\
&= \mathbb{P}\{\tau_{N+n-m,N} \leq n\}, \quad \text{by (4.4)} \\
&= \mathbb{P}\{T_{N+n-m,N} \leq n + N - 1\}, \quad \text{by (4.2).}
\end{aligned}$$

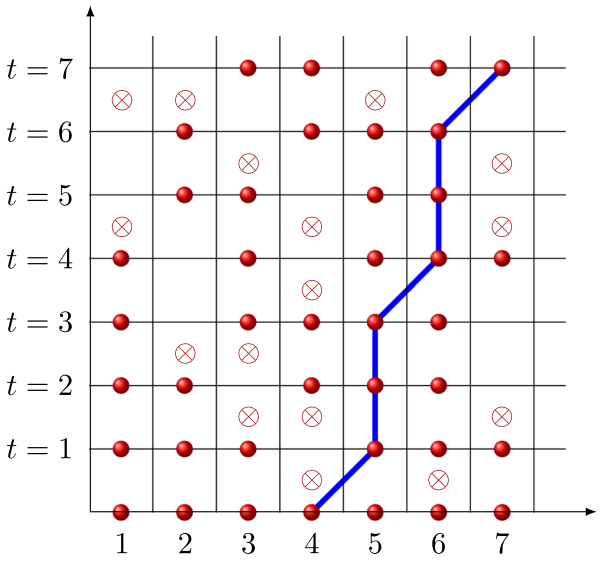

Fig. 7 The DTASEP transformed in the BLIP setting. Symbols ⊗ denote Bernoulli weights 1 to the north-east corner of their square. The coloured balls on each horizontal level is the realization of particles that are still in the 7×7 grid. At $t = 7$ there are 4 particles in the square. From this and (4.3, 4.4) we have that $G^{(0)}_{7,7} = 7 - 4 = 3$

4.1 Proof of Theorem 2.1

Recall that $m_n = n/p - xn^a$ and $a \in \left(0, \frac{1}{2}\right]$. Our goal is to prove that the sequence of random variables $n - G^{(0)}_{n,m_n}$ is tight. The main ingredient in the proof is identity (4.1). Set $N = \frac{nq}{p} - xn^a + k$. Then

$$n - m_n + N = n - \frac{n}{p} + xn^a + \frac{nq}{p} - xn^a + k = k. \tag{4.5}$$

Since $N(n)$ is eventually monotone, we can invert the expression above and find n in terms of N for sufficiently large n (and hence N). In particular,

$$n = n(N) = \frac{p}{q}N + xN^a \left(\frac{p}{q}\right)^{a+1} + O(N^{2a-1}). \tag{4.6}$$

To see this we compute

$$\begin{aligned} N(n(N)) &= \frac{q}{p}n(N) - xn(N)^a + k \\ &= \frac{q}{p}\left(\frac{p}{q}N + xN^a \left(\frac{p}{q}\right)^{a+1} + O(N^{2a-1})\right) \end{aligned}$$

$$
\begin{aligned}
&-x\left(\frac{p}{q}N + xN^a\left(\frac{p}{q}\right)^{a+1} + O(N^{2a-1})\right)^a + k \\
&= N + \left(\frac{p}{q}\right)^a xN^a - x\left(\frac{p}{q}N\right)^a\left(1 + xN^{a-1}\left(\frac{p}{q}\right)^a + O(N^{2a-2})\right)^a + O(1) \\
&= N + \left(\frac{p}{q}\right)^a xN^a - x\left(\frac{p}{q}N\right)^a\left(1 + axN^{a-1}\left(\frac{p}{q}\right)^a + O(N^{2a-2})\right) + O(1) \\
&= N + O(1).
\end{aligned}
$$

Therefore, $n + N - 1 = \frac{N}{q} + x\left(\frac{p}{q}\right)^{a+1} N^a + O(N^{2a-1})$. Combining (4.1) and (4.5)

$$
\begin{aligned}
\mathbb{P}\{k \le n - G^{(0)}_{m_n,n}\} &= \mathbb{P}\left\{G^{(0)}_{m_n,n} \le m_n - N\right\} \\
&= \mathbb{P}\left\{T_{k,N} \le \frac{N}{q} + x\left(\frac{p}{q}\right)^{a+1} N^a + O(N^{2a-1})\right\} \\
&\le \mathbb{P}\left\{\max_{j:1\le j\le k}\sum_{i=1}^{N}\zeta_{i,j} \le \frac{N}{q} + x\left(\frac{p}{q}\right)^{a+1} N^a + O(N^{2a-1})\right\} \\
&= \mathbb{P}\left\{\sum_{i=1}^{N}\zeta_{i,1} - N\mathbb{E}(\zeta_{11}) \le x\left(\frac{p}{q}\right)^{a+1} N^a + O(N^{2a-1})\right\}^k. \quad (4.7)
\end{aligned}
$$

Divide both sides of the inequality inside the probability in (4.7) by $\frac{\sqrt{p}}{q}\sqrt{N}$. The left-hand side of the inequality always converges weakly to a standard Gaussian random variable.

When $a < 1/2$, the right hand side tends to 0 and therefore the probability in (4.7) converges to $1/2$. When $a = 1/2$ the right-hand side in the probability converges to $xpq^{-\frac{1}{2}}$ and the probability in (4.7) to $\Phi(xpq^{-\frac{1}{2}})$. Thus we obtain (2.1).

4.2 Proof of Corollary 2.2

We first show the result when $a < 1/2$. Using (3.12) from Remark 3.6 and (4.7) from the proof of Theorem 2.1, we have

$$
\begin{aligned}
\mathbb{P}\left\{k < R^{(\mathrm{ind})}_{\frac{n}{p}-xn^a} n\right\} &\le \mathbb{P}\left\{k < n - G^{(0)}_{\frac{n}{p}-xn^a,n}\right\} \\
&= \left(\mathbb{P}\left\{\frac{\sum_{i=1}^{N}\zeta_{i,1} - \mathbb{E}(\zeta_{i,1})N}{\sqrt{\mathrm{Var}(\zeta_{i,1})N}} < C_1 N^{a-1/2}\right\}\right)^k \quad (4.8)
\end{aligned}
$$

for C_1 large enough. As in the proof of Theorem 2.1 we have $N = \frac{nq}{p} - xn^a + k$ and let Φ denote the cumulative distribution function of the standard normal distribution. Fix a tolerance $\delta > 0$ satisfying $\Phi(\delta) + \delta < 1$ and let $n_1(\delta)$ large enough so that

$C_1 N^{a-1/2} < \delta$ for all $n > n_1(\delta)$. Applying the Berry-Esseen theorem to the last line of the last display,

$$\mathbb{P}\{k \le R^{(\mathrm{ind})}_{\frac{n}{p}-xn^a} n\} \le \left(\Phi(\delta) + \frac{C}{\sqrt{n}}\right)^k \le (\Phi(\delta)+\delta)^k, \quad \text{for all } n > n_2(\delta). \tag{4.9}$$

For $n \ge n_0(\delta) = n_1(\delta) \vee n_2(\delta)$ the right hand side of (4.9) is uniformly summable in k. Moreover, by (4.9) and the reverse Fatou's Lemma we compute

$$\begin{aligned}\overline{\lim_{n\to\infty}} \mathbb{E}\left[R^{(\mathrm{ind})}_{\frac{n}{p}-xn^a} n\right] &= \overline{\lim_{n\to\infty}} \sum_{k=0}^{\infty} \mathbb{P}\left\{k \le R^{(\mathrm{ind})}_{\frac{n}{p}-xn^a} n\right\} \\ &\le \sum_{k=0}^{\infty} \overline{\lim_{n\to\infty}} \mathbb{P}\left\{k < n - G^{(0)}_{\frac{n}{p}-xn^a,n}\right\} \le \sum_{k=0}^{\infty} 2^{-k} = 2,\end{aligned}$$

where the penultimate inequality follows from (3.12) and the last from Theorem 2.1.

The case $a = \frac{1}{2}$ is slightly more delicate, but the ideas are exactly the same. As before,

$$\mathbb{P}\left\{k < R^{(\mathrm{ind})}_{\frac{n}{p}-x\sqrt{n}} n\right\} \le \left(\mathbb{P}\left\{\frac{\sum_{i=1}^N \zeta_i - \mathbb{E}(\zeta_1)N}{\sqrt{\mathrm{Var}(\zeta_1)N}} < x\frac{p}{\sqrt{q}} + C_0 N^{-1/2}\right\}\right)^k. \tag{4.10}$$

The right-hand side converges to $(\Phi(xpq^{-\frac{1}{2}}))^k$ and with the same arguments as before,

$$\overline{\lim_{n\to\infty}} \mathbb{E}\left[R^{(\mathrm{ind})}_{\frac{n}{p}-xn^a} n\right] \le \frac{1}{1-\Phi(xpq^{-\frac{1}{2}})}.$$

4.3 Proof of Theorem 2.3

When $a \le 1/2$ the result follows from (4.8) and (4.10). For $a \in (\frac{1}{2}, \frac{3}{4}]$

$$\begin{aligned}&\overline{\lim_{n\to\infty}} \mathbb{P}\left\{\left(\frac{(px)^2}{4(1-p)} + \varepsilon\right) n^{2a-1} \le R^{(\mathrm{ind})}_{p^{-1}n-xn^a} n\right\} \\ &\le \overline{\lim_{n\to\infty}} \mathbb{P}\left\{\left(\frac{(px)^2}{4(1-p)} + \varepsilon\right) n^{2a-1} \le n - G^{(0)}_{p^{-1}n-xn^a,n}\right\} = 0.\end{aligned}$$

The last inequality follows from (3.12) and the last equality is from (1.4). This gives the second part of the statement.

When $a \in \left(\frac{3}{4}, 1\right)$ we can obtain a sharper bound using Lemma 3.7.

From the proof of Lemmas 3.5 and 3.9 we can find a constant C_1 such that $n - G^{(\beta_R)}_{p^{-1}n-xn^a,n} = n - z_R < C_1 n^a$ in probability, as n grows. Therefore, with probability tending to 1 as n grows,

$$z_0 - z_R < C_1 n^a. \tag{4.11}$$

Moreover, since the the number of vertical steps at $\beta = 0$ cannot exceed $n - G^{(0)}_{p^{-1}n-xn^a,n}$, (1.4) gives that with probability tending to 1

$$y_0 - y_R \le n - G^{(0)}_{p^{-1}n-xn^a,n} < C_2 n^{2a-1}. \tag{4.12}$$

Equations (4.11) and (4.12) now yield a constant C such that

$$\lim_{n\to\infty} \mathbb{P}\{z_0 - z_R + y_0 - y_R < Cn^a\} = 1. \tag{4.13}$$

Let A_n be the event in the probability above. On A_n, $\sum_{i=0}^{R-1}\{(z_i - z_{i+1}) + (y_i - y_{i+1})\} < Cn^a$. Now we are in a position to use Lemma 3.7 and finish the proof.

4.4 Proof of Theorem 2.4 (Edge Fluctuations for the Independent Model)

We will once more use (4.1). Recall that

$$x = \frac{2}{\sqrt{p}}\left(\frac{q}{p}\right)^a \quad \text{and} \quad y = s\frac{\sqrt{p}}{q}\left(\frac{p}{q}\right)^{\frac{1+a}{3}}, \quad s \in \mathbb{R}.$$

We further define an auxiliary parameter N that will go to ∞ when n goes to infinity.

$$N = N(n) = \frac{q}{p}n - xn^a - yn^{\frac{2-a}{3}} + c_n, \tag{4.14}$$

where c_n is given by

$$c_n = \begin{cases} \left(\frac{q}{p}\right)^{2a-1} n^{2a-1}, & 1/2 < a < 2/3, \\ \left(\frac{q}{p}\right)^{2a-1} n^{2a-1} - (2a-1)x\left(\frac{q}{p}\right)^{2a-2} n^{3a-2}, & 2/3 \le a < 5/7. \end{cases} \tag{4.15}$$

Note that with $m_n = \frac{1}{p}n - xn^a - yn^{\frac{2-a}{3}}$ we have the relation

$$m_n - n = N - c_n. \tag{4.16}$$

Our goal now is to change n to N and compute m_n, n, c_n in terms of N, similarly to the proof of Theorem 2.1.

(1) **Step 1: $m_n - n$ and c_n as a function of N:** Start from (4.14) and raise it to the power $2a-1$. Then, apply Taylor's theorem to obtain

$$N^{2a-1} = \left(\frac{q}{p}n\right)^{2a-1}\left(1-(2a-1)\frac{px}{q}n^{a-1} + O\left(n^{-\frac{1+a}{3}}\right)\right) = c_n + O(n^{\frac{5a-4}{3}}).$$

Note that the equation above holds, irrespective of the value of a, as long as $a < 5/7$; for $a \in [0, \frac{5}{7})$ the exponent $\frac{5a-4}{3} < 0$, so

$$c_n = N^{2a-1} + o(1)$$

follows. Therefore, a substitution in (4.16) yields

$$m_n - n = N - N^{2a-1} + o(1). \tag{4.17}$$

(2) **Step 2: n as a function of N**: We begin by writing n as a function of N. Observe that $N(n)$ in (4.14) is an eventually monotone function. Therefore, for N large enough, there is a well defined inverse $n = n(N)$ (so that $N(n(N)) = N$).

We cannot directly use a closed formula for the inverse, so we define the approximate inverse $\ell(N)$ by

$$\ell(N) = \frac{p}{q}N + \frac{2\sqrt{p}}{q}N^a + y\left(\frac{q}{p}\right)^{\frac{1+a}{3}} N^{\frac{2-a}{3}}.$$

To see that $\ell(N)$ plays the role of the inverse $n(N)$, substitute $\ell(N)$ in (4.14) and estimate using a Taylor expansion the distance

$$|N - N(\ell(N))| = |N - \frac{q}{p}\ell(N) + x_{p,a}\ell(N)^a + y\ell(N)^{\frac{2-a}{3}}| = O(N^{2a-1}). \quad (4.18)$$

This implies that $|n(N) - \ell(N)| = o(N^{\frac{2-a}{3}})$; in fact we will show that =

$$|n(N) - \ell(N)| < cN^{\beta}, \quad (4.19)$$

for any $\beta \in \left(2a-1, \frac{2-a}{3}\right)$. Assume for a contradiction that (4.19) does not hold for some $c > 0$ and for some $\beta > 2a - 1$. Then

$$\begin{aligned}
|N - N(\ell(N))| &= |N(n(N)) - N(\ell(N))| \\
&= \Big| \frac{q}{p}(n(N) - \ell(N)) - x(n(N)^a - \ell(N)^a) \\
&\quad - y(n(N)^{\frac{2-a}{3}} - \ell(N)^{\frac{2-a}{3}}) + c_{n(N)} - c_{\ell(N)}) \Big| \\
&\geq \frac{q}{p}|n(N) - \ell(N)| - x|n(N) - \ell(N)|^a - |y||n(N) - \ell(N)|^{\frac{2-a}{3}} \\
&\quad - |c_{n(N)} - c_{\ell(N)}| \\
&\geq CN^{\beta} \text{ for some } C > 0 \text{ and } N \text{ large enough.}
\end{aligned}$$

This contradicts (4.18) since $\beta > 2a - 1$. In particular we have shown that

$$\lim_{N\to\infty} \frac{|n(N) - \ell(N)|}{N^{\frac{2-a}{3}}} = \lim_{N\to\infty} \frac{n(N) - \frac{p}{q}N - \frac{2\sqrt{p}}{q}N^a - y\left(\frac{q}{p}\right)^{\frac{1+a}{3}} N^{\frac{2-a}{3}}}{N^{\frac{2-a}{3}}} = 0, \quad (4.20)$$

and we may write

$$n = \frac{p}{q}N + \frac{2\sqrt{p}}{q}N^a + y\left(\frac{q}{p}\right)^{\frac{1+a}{3}} N^{\frac{2-a}{3}} + o(N^{\frac{2-a}{3}}) = \ell(N) + o(N^{\frac{2-a}{3}}). \quad (4.21)$$

To finish the proof we need to be a bit cautious with the integer parts. Define k_N to be

$$k_N = \lfloor m_n \rfloor - n - \lfloor N \rfloor + \left\lfloor \lfloor N \rfloor^{2a-1} \right\rfloor.$$

It follows from (4.17) that k_N is bounded in N (and n). Also set $N = \lfloor N \rfloor + \varepsilon_N$ so that $\varepsilon_N \in [0, 1)$. Substituting these in (4.1) we compute

$$\begin{aligned}
&\mathbb{P}\left\{ G^{(0)}_{\lfloor m_n \rfloor, n} \leq n - \left\lfloor \lfloor N \rfloor^{2a-1} \right\rfloor \right\} = \mathbb{P}\{T_{\lfloor \lfloor N \rfloor^{2a-1} \rfloor, \lfloor N \rfloor + k_N} \leq n + \lfloor N \rfloor - 1\} \\
&= \mathbb{P}\left\{ T_{\lfloor \lfloor N \rfloor^{2a-1} \rfloor, \lfloor N \rfloor + k_N} \leq \ell(N) + o\left(N^{\frac{2-a}{3}}\right) + N - 1 + \varepsilon_N \right\}
\end{aligned}$$

$$= \mathbb{P}\left\{ T_{\lfloor \lfloor N\rfloor^{2a-1}\rfloor, \lfloor N\rfloor + k_N} \le \frac{p}{q}N + \frac{2\sqrt{p}}{q}N^a + y\left(\frac{q}{p}\right)^{\frac{1+a}{3}} N^{\frac{2-a}{3}} + N + o\left(N^{\frac{2-a}{3}}\right)\right\}$$

$$= \mathbb{P}\left\{ \frac{T_{\lfloor \lfloor N\rfloor^{2a-1}\rfloor, \lfloor N\rfloor + k_N} - \frac{1}{q}\lfloor N\rfloor - \frac{2\sqrt{p}}{q}\lfloor N\rfloor^a}{\frac{\sqrt{p}}{q}\lfloor N\rfloor^{\frac{2-a}{3}}} \le s + o(1)\right\}. \tag{4.22}$$

The passage time in the probability above can be compared with $T_{\lfloor N^{2a-1}\rfloor, \lfloor N\rfloor}$ and satisfies

$$|T_{\lfloor \lfloor N\rfloor^{2a-1}\rfloor, \lfloor N\rfloor} - T_{\lfloor \lfloor N\rfloor^{2a-1}\rfloor, \lfloor N\rfloor + k_N}| < \sum_{i=0}^{\lfloor \lfloor N\rfloor^{2a-1}\rfloor} \sum_{j=-k_N}^{k_N} \zeta_{i, \lfloor N\rfloor + j}.$$

Since $a < \frac{5}{7}$, the number of geometric random variables in the right-hand side of the inequality is of lower order than $N^{\frac{2-a}{3}}$ and when scaled by it, the double sum vanishes $\mathbb{P}$-a.s. This allows us to remove k_N from (4.22) and (1.5) now gives the result by taking $n \to \infty$.

5 Optimality Regions in the Alignment Model

In this section we prove our results about the alignment model. Because of Lemma 3.3 and (3.7) it is enough to consider the case where $\alpha = 0$.

Now it is straight-forward to prove Theorems 2.6 and 2.7.

5.1 Proof of Theorem 2.6

Restrict to the full measure set of environments so that Lemma 3.9 is in effect. Fix one such environment and assume n is large enough so that statements (1)-(3) of Lemma 3.9 hold. Let

$$g^{(a)}(n) = \begin{cases} \sqrt{n \log n}, & a \le 1/2, \\ n^a, & a > 1/2. \end{cases}$$

Path π_S is admissible under any penalty β, therefore by part (c) of Lemma 3.9,

$$L^{(\beta)}_{\lfloor n|\mathcal{A}| - xn^a\rfloor, n} \ge n - u_n(a)\cdot e_2 + \sum_{k=1}^{u_n(a)\cdot e_2} \omega_{k,k} - \beta((|\mathcal{A}| - 1)n - xn^a).$$

Re-arranging the terms we obtain

$$u_n(a)\cdot e_2 - \sum_{k=1}^{u_n(a)\cdot e_2} \omega_{k,k} - \beta x n^a \ge n(1 + \beta - \beta|\mathcal{A}|) - L^{(\beta)}_{\lfloor n|\mathcal{A}| - xn^a\rfloor, n}.$$

Now divide both sides by $g^{(a)}(n)$ and take the $\overline{\lim}$ as $n \to \infty$ to obtain

$$\overline{\lim_{n\to\infty}} \frac{n(1+\beta-\beta|\mathcal{A}|) - L^{(\beta)}_{\lfloor n|\mathcal{A}|-xn^a\rfloor,n}}{g^{(a)}(n)} \le \begin{cases} \sqrt{c_1} - \frac{1}{|\mathcal{A}|}, & a \le 1/2, \\ \frac{1}{|\mathcal{A}|(|\mathcal{A}|-1)} - \beta x, & a > 1/2. \end{cases} \tag{5.1}$$

Let $c_1 \searrow \frac{2}{(|\mathcal{A}|-1)^2}$ to obtain the upper bound in the theorem.

For the lower bound, recall that the maximum possible positive weight for $L^{(\beta)}_{\lfloor n|\mathcal{A}|-xn^a\rfloor,n}$ is n and the smallest possible gap penalty is $\beta(\lfloor n|\mathcal{A}| - xn^a\rfloor - n)$. Therefore

$$\lim_{n\to\infty} \frac{n(1+\beta-\beta|\mathcal{A}|) - L^{(\beta)}_{\lfloor n|\mathcal{A}|-xn^a\rfloor,n}}{g^{(a)}(n)} \ge \begin{cases} 0, & a \le 1/2, \\ -\beta x, & a > 1/2. \end{cases}$$

This completes the proof. □

5.2 Proof of Theorem 2.7

From the previous theorem, we have that for $\beta = 0$, for $\mathbb{P}$-a.e.ω and any $\varepsilon > 0$, we can find an $N = N(\omega, \varepsilon)$ so that for all $n > N$

$$n \ge L^{(0)}_{\lfloor n|\mathcal{A}|-xn^a\rfloor,n} \ge n - (C(x,|\mathcal{A}|) + \varepsilon)g^{(a)}(n).$$

From (3.2) and the equation above, we immediately obtain, by setting $x_0 = 0$, that

$$z_0 \le n, \quad y_0 \le 2(C(x,|\mathcal{A}|) + \varepsilon)g^{(a)}(n) + \left\lfloor |\mathcal{A}|n - xn^a \right\rfloor - n. \tag{5.2}$$

We briefly explain the upper bound for y_0. First, any maximal path will always take the minimum number of gaps, which is $\lfloor |\mathcal{A}|n - xn^a\rfloor - n$. After that, it has to take the correct number of diagonal steps to gain weight equal to $L^{(0)}_{\lfloor n|\mathcal{A}|-xn^a\rfloor,n}$. Now all the remaining steps can either be gaps or mismatches, so we obtain an upper bound if we assume the number of mismatches is zero. The bound then follows from (3.2).

Similarly, for $\beta = \beta_R$, since π_S can be optimal under this penalty, Lemma 3.9 implies

$$z_R \ge n - u_n(a)\cdot e_2 + \sum_{k=1}^{u_n(a)\cdot e_2} \omega_{k,k} \ge n - u_n(a)\cdot e_2, \quad \text{and} y_R = \left\lfloor |\mathcal{A}|n - xn^a \right\rfloor - n. \tag{5.3}$$

Combine (5.2) and (5.3) to obtain for some uniform constant C

$$z_0 - z_R + y_0 - y_R \le u_n(a)\cdot e_2 + 2(C(x,|\mathcal{A}|) + \varepsilon)g^{(a)}(n) \le Cg^{(a)}(n),$$

and the result follows from Lemma 3.7. □

5.3 Proof of Theorem 2.8

Lemma 3.9-(3) implies that if $\bar{\pi}_S$ exits from the north boundary,

$$z_\beta(\bar{\pi}_S) \ge n - u_n(a)\cdot e_2 + \sum_{k=1}^{u_n(a)\cdot e_2} \omega_{k,k} \ge n - u_n(a)\cdot e_2, \text{ for all } \beta > 0. \tag{5.4}$$

Let B_n denote the event (5.4) and D_n the event that $\bar{\pi}_S$ exits from the north boundary. Choose $c_1 = c_2 = 12/(|\mathcal{A}| - 1)^2$ in the definition of $u_n(a)$ in (3.16). Then it follows from (3.21) and (3.22), using Lemma 3.8, that $\bar{\pi}_S$ exits from the north boundary with probability at least $1 - c_0(n^3 \log n)^{-1}$. Now, since $z_0 \le n$,

$$D_n \subseteq B_n \subseteq \{z_0 - z_R \le u_n(a)\}. \tag{5.5}$$

On the other hand, since $z_0 \ge n - u_n(a) \cdot e_2$, (3.2) implies that

$$y_0 \le 2u_n(a) \cdot e_2 + \lfloor |\mathcal{A}|n - xn^a \rfloor - n = 2u_n(a) \cdot e_2 - y_R.$$

Therefore

$$D_n \subseteq \{y_0 - y_R \le 2u_n(a) \cdot e_2\}. \tag{5.6}$$

Combine (5.5) and (5.6) to deduce

$$D_n \subseteq \{y_0 - y_R \le 2u_n(a) \cdot e_2\} \cap \{z_0 - z_R \le u_n(a) \cdot e_2\} \subseteq \{z_0 - z_R + y_0 - y_R \le 3u_n(a) \cdot e_2\}.$$

Finally, use (3.13) to obtain that for all $n > N = N(a, x)$,

$$D_n \subseteq \{R_{m,n} \le C(u_n(a) \cdot e_2)^{2/3}\}. \tag{5.7}$$

On the complement of D_n we bound R by n, by virtue of (3.12). Then for n large enough,

$$\begin{aligned}
\mathbb{E}\Big(R^{(\mathrm{al})}_{\lfloor n|\mathcal{A}| - xn^a \rfloor} n\Big) &\le \mathbb{E}(R^{(\mathrm{al})}_{\lfloor n|\mathcal{A}| - xn^a \rfloor} n \mathbb{1}\{D_n\}) + n\mathbb{P}\{D_n^c\} \\
&\le \mathbb{E}(R^{(\mathrm{al})}_{\lfloor n|\mathcal{A}| - xn^a \rfloor} n \mathbb{1}\{R^{(\mathrm{al})}_{\lfloor n|\mathcal{A}| - xn^a \rfloor} n \le C(u_n(a) \cdot e_2)^{2/3}\}) + n\mathbb{P}\left\{D_n^c\right\} \\
&\le \begin{cases} C(x, |\mathcal{A}|)(n \log n)^{1/3} & a \le 1/2, \\ C(x, |\mathcal{A}|)n^{2a/3}, & a > 1/2. \end{cases}
\end{aligned}$$

This gives the result. □

Acknowledgments We thank the anonymous referees for their helpful comments, which have led to a much improved version of the article. We also thank Lior Pachter for pointing out the reference [45] to us.

References

1. Aluru, S.: Handbook of computational molecular biology. Chapman & Hall/CRC Computer and Information Science Series. Chapman & Hall/CRC, Boca Raton (2006)
2. Amsalu, S., Matzinger, H., Vachkovskaia, M.: Thermodynamical approach to the longest common subsequence problem. J. Stat. Phys. **131**(6), 1103–1120 (2008)
3. Apostol, T.M.: Introduction to analytic number theory, 5th edition ed. Undergraduate Texts in Mathematics. Springer (1995)
4. Baryshnikov, Y.: GUEs queues. Prob. Theory Relat. Fields **119**, 256–274 (2001)
5. Basdevant, A.-L., Enriquez, N., Gerin, L., Gouéré, J.-B.: Discrete Hammersley's lines with sources and sinks. ALEA Lat. Am. J. Probab. Math. Stat **13**, 33–52 (2016)
6. Bergroth, L., Hakonen, H., Raita, T.: A survey of longest common subsequence algorithms. SPIRE **00**, 39–48 (2000)

7. Bodineau, T., Martin, J.: A universality property for last-passage percolation close to the axis. Electron. Commun. Probab. **10**(11), 105–112 (2005)
8. Chvátal, V., Sankoff, D.: Longest common subsequences of two random sequences. J. Appl. Probab. **12**(2), 306–315 (1975)
9. Cramèr, H.: Sur un nouveau théorème limite de la probabilité. Actualités Sci. Industr **736**, 5–23 (1938)
10. Dewey, C.N., Huggins, P.M., Woods, K., Sturmfels, B., Pachter, L.: Parametric alignment of drosophila genomes. PLoS Comput. Biol. **2**(6), e73 (2006)
11. Fernández-Baca, D., Seppäläinen, T., Slutzki, G.: Bounds for parametric sequence comparison. Discrete Appl. Math **118**, 181–198 (2002)
12. Fernández-Baca, D., Venkatachalam, B.: Parametric sequence alignment. CRC Press Computer and Information Science Series. Chapman and Hall (2006)
13. Georgiou, N.: Soft edge results for longest increasing paths on the planar lattice. Electron. J. Probab. **15**, 1–13 (2010)
14. Georgiou, N., Rassoul-Agha, F., Seppäläinen, T.: Variational formulas and cocycle solutions for directed polymer and percolation models. Commun. Math. Phys. **346**(2), 741–779 (2016)
15. Georgiou, N., Rassoul-Agha F., Seppäläinen, T.: Stationary cocycles and Busemann functions for the corner growth model. Probab. Theory Relat. Fields. https://doi.org/10.1007/s00440-016-0729-x (2016)
16. Georgiou, N., Rassoul-Agha, F., Seppäläinen, T.: Geodesics and the competition interface for the corner growth model. Probab. Theory Relat. Fields. https://doi.org/10.1007/s00440-016-0734-0 (2016)
17. Glynn, P.W., Whitt, W.: Departures from many queues in series. Ann. Appl. Probab **1**(4), 546–572 (1991)
18. Gong, R., Houdré, C., Lember, J.: Lower bounds on the generalized central moments of the optimal alignments score of random sequences. arXiv:1506.06067 (2015)
19. Gusfield, D., Balasubramanian, K., Naor, D.: Parametric optimization of sequence alignment. Algorithmica **12**(4-5), 312–326 (1994)
20. Hammersley, J.M.: A few seedlings of research. In: Proceedings of the 6th berkeley symposium on mathematical statistics and probability (University California, Berkeley, Calif., 1970/1971), Vol. I: Theory of statistics, pp. 345–394. University California Press, Berkeley (1972)
21. Henikoff, S., Henikoff, J.G.: Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA **89**(22), 10915–10919 (1992)
22. Hirschberg, D.S.: A linear space algorithm for computing maximal common subsequences. Commun. ACM **18**(6), 341–343 (1975)
23. Houdré, C., Matzinger, H.: Closeness to the diagonal for longest common subsequences in random words. Electron. Commun. Probab. **21**(36), 1–19 (2016)
24. Hower, V., Heitsch, C.E.: Parametric analysis of RNA branching configurations. Bull. Math. Biol. **73**(4), 754–776 (2011)
25. Kiwi, M., Loebl, M., Matoušek, J.: Expected length of the longest common subsequence for large alphabets. Adv. Math **197**(2), 480–498 (2005)
26. Komlós, J., Major, P., Tusnády, G.: An approximation of partial sums of independent RV's, and the sample DF. II. Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **34**, 33–58 (1976)
27. Lember, J., Matzinger, H.: Standard deviation of the longest common subsequence. Ann. Probab. **37**(3), 1192–1235 (2009)
28. Lember, J., Matzinger, H., Vollmer, A.: Optimal alignments of longest common subsequences and their path properties. Bernoulli **20**(3), 1292–1343 (2014)
29. Maier, D.: The complexity of some problems on subsequences and supersequences. J. ACM **25**(2), 322–336 (1987)
30. Malaspinas, A.S., Eriksson, N., Huggins, P.M.: Parametric analysis of alignment and phylogenetic uncertainty. Bull. Math. Biol. **73**(4), 795–810 (2011)
31. Martin, J.B.: Limiting shape for directed percolation models. Ann. Probab. **32**(4), 2908–2937 (2004)
32. Masek, W.J., Paterson, M.S.: A faster algorithm computing string edit distances. J. Comput. Syst. Sci. **20**(1), 18–31 (1980)
33. Myers, E.W., Miller, W.: Optimal alignments in linear space. Comput. Appl. Biosci. **4**(1), 11–17 (1988)
34. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol **48**, 443–453 (1970)

35. Ng, P.C., Henikoff, S.: Predicting deleterious amino acid substitutions. Genome Res. **11**(5), 863–874 (2001)
36. O'Connell, N., Yor, M.: Brownian analogues of Burke's theorem. Stoch. Proc. Appl. **96**, 285–304 (2001)
37. Pachter, L., Sturmfels, B.: Parametric inference for biological sequence analysis. Proc. Natl. Acad. Sci. U.S.A. **101**(46), 16138–16143 (2004)
38. Pachter, L., Sturmfels, B.: Algebraic statistics for computational biology. Cambridge University Press, New York (2005)
39. Priezzev, V.B., Schütz, G.M.: Exact solution of the Bernoulli matching model of sequence alignment. J. Stat. Mech. Theor. Exp. **2008**(09), P09007 (2008)
40. Seppäläinen, T.: Increasing sequences of independent points on the planar lattice. Ann. Appl. Probab. **7**(4), 886–898 (1997)
41. Seppäläinen, T.: A scaling limit for queues in series. Ann. Appl. Probab. **7**(4), 855–872 (1997)
42. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. J. Mol. Biol. **147**, 195–197 (1981)
43. Tracy, C.A., Widom, H.: Level-spacing distributions and the airy kernel. Comm. Math. Phys. **159**(1), 151–174 (1994)
44. Vingron, M., Waterman, M.S.: Sequence alignment and penalty choice. review of concepts, case studies and implications. J. Mol. Biol **235**(1), 1–12 (1994)
45. Vinzant, C.: Lower bounds for optimal alignments of binary sequences. Discret. Appl. Math. **157**, 3341–3346 (2009)
46. Xia, X.: Bioinformatics and the cell: Modern computational approaches in genomics, Proteomics and transcriptomics. Springer, Berlin (2007)